Analysis and Detection of Spamming Behavior in Blog Networks

Enhua Tan[1], Lei Guo[2], Songqing Chen[3], Xiaodong Zhang[1]
[1]The Ohio State University [2]Yahoo! Inc. [3]George Mason University

Spam in UGC based social networks

- Spam in UGC (User Generated Content) based online social networks (OSN), such as blogs, has increased explosively on the Internet in recent years.
  - Spam repeatedly posted to a blog system to pollute top article list and the most recent article list
  - Spam using popular keywords to rank high in search results
  - Estimated that 75% active posts in Google Blogspot are spam
- Spamming in UGC social networks is easy to conduct but hard to prevent comparing with email spam:
  - Easier to collect spamming targets by identifying UGC sites
  - Easier to post a spam article than to send a spam email
    - No systematic prevention like email server authentication
    - CAPTCHA is only for user registration instead of posting
  - Lacks legal restrictions on spamming in UGC sites

Problem Statement

- Content-based spam detection limitations for blog spam filtering:
  - New labeling data is not constantly available as in email system
  - Spam copied from recent web articles are hard to be classified
- **Research Problem:**
  - Understanding the inherent patterns of blog spamming behavior
  - Effectively detect blog spam in adapting to Internet environment

Blog Measurement & Analysis

- Analyze the database dump of Yahoo! Hong Kong (HK) Blog
  - a large blog site in Asia
  - over **6 million posts** and nearly 400 thousand users (Table 1)
  - 325 days (from October 2008 to August 2009)
- Constant posting pattern of spammers (except for midnight) (Fig. 1)
  - spammers are paid for posting
- Spam consumes up to 84.2% of posting upload bandwidth (Fig. 2)
  - overwhelming resource consumption of spammers
- Spammers post more frequently than other users, with constant frequency (Fig. 4)
  - using bots-like automatic tools
- Most of the posts by spammers contain links to customers (Fig. 6)
  - avoid text-only posts that cannot get any clicks

Table 1: Summary of Yahoo! HK Blog Workload

Type	Number	% of Total
posts	6,595,917	100
posts w/ link	1,832,112	27.8
spam posts	1,371,968	20.8
xyz spam posts	1,130,718	17.1
users	382,090	100
users w/ link	157,305	41.2
spammers	79,490	20.8
xyz spammers	12,116	3.2

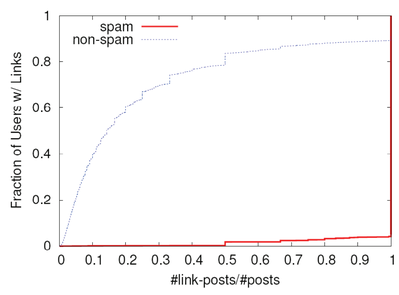
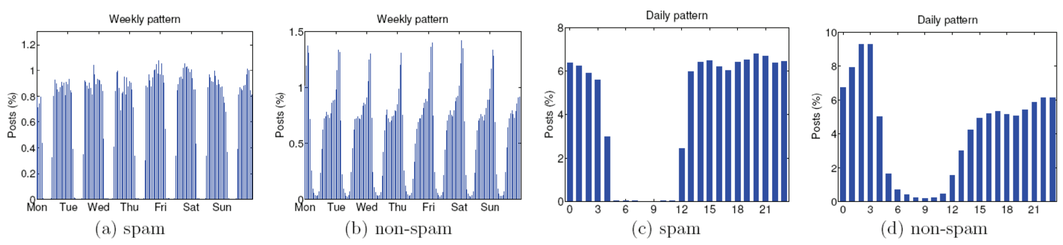
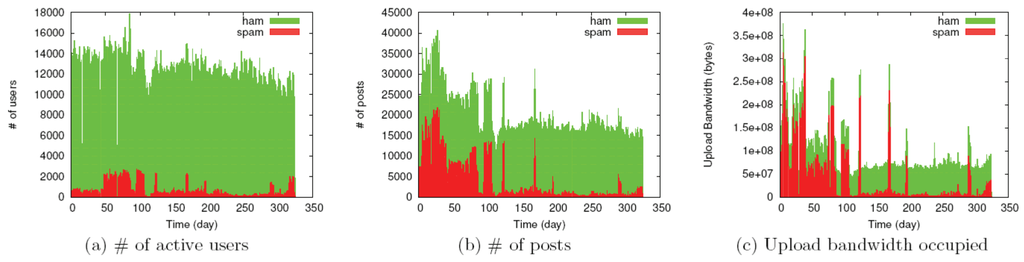


Figure 6: Ratio of # of posts with links per user



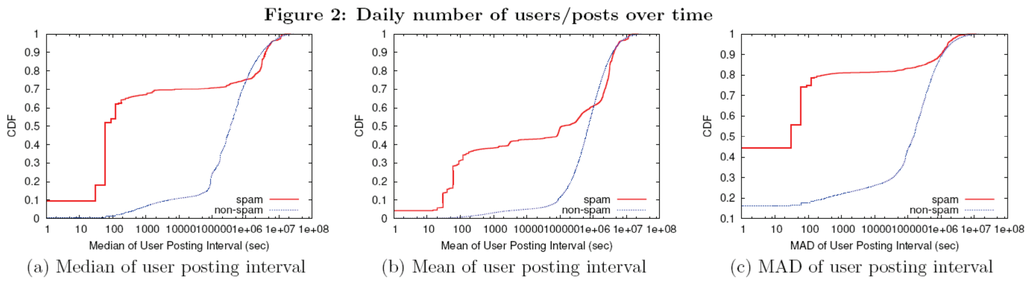
(a) spam (b) non-spam (c) spam (d) non-spam

Figure 1: Weekly and daily pattern of spam and non-spam blog posts



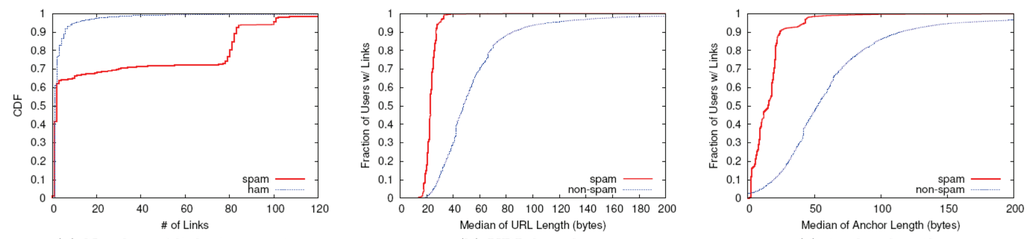
(a) # of active users (b) # of posts (c) Upload bandwidth occupied

Figure 2: Daily number of users/posts over time



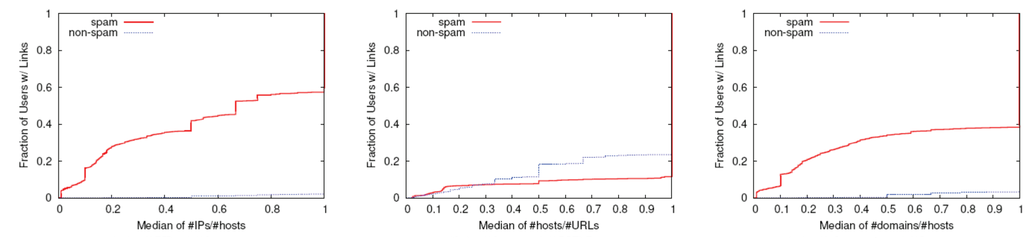
(a) Median of user posting interval (b) Mean of user posting interval (c) MAD of user posting interval

Figure 4: User posting interval



(a) Number of links in a post (b) URL length (c) Anchor length

Figure 7: Link related Properties



(a) #IPs/#hosts (b) #hosts/#URLs (c) #domains/#hosts

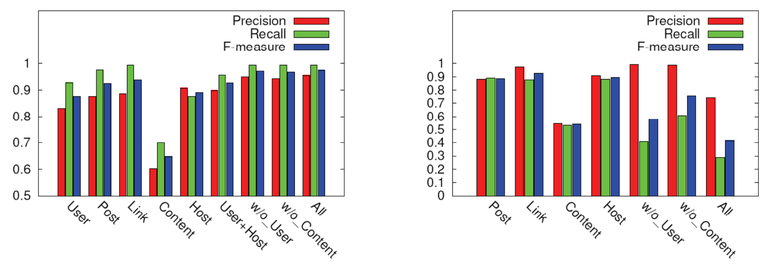
Figure 9: Host related Properties

Table 3: List of features and feature sets

Feature sets	Features
User activity	median of posting interval MAD of posting interval active duration post-in-midnight
Post contributions	#link-posts #text-posts #link-posts/#posts
Link properties	#URLs MAD of #URLs URL length anchor length
Host properties	#hosts MAD of #hosts #IPs/#hosts MAD of #IPs/#hosts #domains MAD of #domains #domains/#hosts #hosts/#URLs
Content Metadata	content length content entropy

Table 4: Classification results of different methods on labeled data set

Methods	Precision	Recall	F-measure
SVM(RBF)	0.742	0.290	0.417
Naive Bayes	0.951	0.992	0.971
Logistic Regression	0.955	0.995	0.975
Decision Tree	0.975	0.989	0.982



(a) LR classification (b) SVM classification

Figure 11: classification results using different feature sets

- Spammers tend to post either 1 or 2 links in their posts like most normal users, or more than 70 links (Fig. 7a)
  - post more links to boost more spam sites with least effort
- More than half of the spammers' posts point to hosts with less IPs while normal bloggers' posts are mostly pointing to hosts with different IPs (Fig. 9a)
  - due to extra cost to obtain individual IP addresses for spammers
- Normal blogger tends to post > 1 URL for each host, and spammers prefer > 1 host for each domain (Fig. 9b,c)
  - Avoid spam host blacklisted by blog admin

Blog Spam Detection

- Table 3 shows five sets of features summarized based on measurement analysis for spam classification
- Classifiers are built based on Orange:
  - Naive Bayes, Support Vector Machine (SVM) with RBF kernel, Logistic Regression (LR), and Decision Tree (C4.5)

Spam Detection Results

- Labeled 2167 users with links in posts
- All experiments are performed by using 10-fold cross-validation to avoid biased selections of training and testing sets
- Table 4 shows the classification results by using all the features:
  - Decision Tree has the highest precision rate of 97.5% (with 98.9% recall)
  - LR has the highest recall rate of 99.5% (with 95.5% precision)
- Figure 11(a) shows the performance of each feature set by using the LR classifier:
  - content metadata features can only provide preliminary performance
  - link properties show the best over-all performance
  - "User+Host" shows that combining two feature sets can improve classification results
  - "w/o Content" shows that removing content metadata information only slightly degrades the overall performance (1% on the precision rate)

Conclusion

- Spam posting is always in a productive manner as spamming is a professional job nowadays.
- Spam site links in spam posts have much smaller URL length and anchor length due to spammers' advertising purposes.
- Spam hosts redirected by spam posts often share IP addresses due to the cost-effectiveness considerations by spammers.
- Our spam detection method can achieve precision and recall rates of more than 95% without the help of content-based classifications